DOCUMENT RESUME

ED 211 592 TM 820 025

AUTHOR	McArthur, David
TITLE	Test Design Project: Studies in Test Bias. Annual Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
PUB DATE	1 Nov 81
GRANT	NIE-G-80-0112
NOTE	87p.; For related documents see TM 820 024 and TM 820 026.
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Bilingual Education; Bilingual Students; Elementary Education; *Ethnicity; Non English Speaking; *Research Methodology; *Statistical Analysis; *Test Bias; Test Construction; Writing Evaluation
IDENTIFIERS	Comprehensive Tests of Basic Skills

ABSTRACT

Item bias in a multiple-choice test can be detected by appropriate analyses of the persons x items scoring matrix. This permits comparison of groups of examinees tested with the same instrument. The test may be biased if it is not measuring the same thing in comparable groups, if groups are responding to different aspects of the test items, or if cultural and linguistic issues take precedence. An empirical study of the question of bias as shown by these techniques was conducted. Five related schemes for the statistical analysis of bias were applied to the Comprehensive Test of Basic Skills which was administered in either the English or Spanish language version at two levels of elementary school in bilingual education programs. The objectives measured were recall or recognition ability, ability to translate or convert verbal or symbolic concepts, ability to comprehend concepts, ability to apply techniques, and ability to extend interpretation beyond stated information. The results indicated that several items in the tests showed strong evidence of bias, corroborated by a separate analysis of linguistic and cultural sources of bias for many items. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

e Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.C. Beer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DELIVERABLE - November 1, 1981

TEST DESIGN PROJECT: STUDIES IN TEST BIAS

Annual Report

David McArthur, Study Director

Grant Number
NIE-G-80-0112

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

TABLE OF CONTENTS

Introduction to the Test Design Project Studies in Test Bias
D. McArthur

Detection of item bias using analyses of response patterns (Appendix A)
D. McArthur

Performance patterns of bilingual children tested in both languages (Appendix B) D. McArthur

Bias in the writing of prose and its appraisal (Appendix C)
D. McArthur

Potential sources of bias in dual language achievement tests (Appendix D) B. Cabello

Cultural interference in reading comprehension: An alternative explanation (Appendix E) B. Cabello

Introduction to the Test Design Project Studies in Test Bias

The assessment of literacy in bilingual and limited English proficient (LEP) students is a distinct problem area in applied psychometrics. Non-native speaking students, some with generally impoverished language skills, others simply weak in English, constitute a substantial proportion of the student population in many regions of the United States. The instruction and assessment of these students is an issue of national concern. Even after placement into monolingual English classes, cultural group characteristics continue to interact with the instruction these students receive and to influence their performance on tests. Lower levels of performance by bilingual and LEP students, whether from a test given in English or given in their native language, may come about either because schools cannot provide appropriate instruction or because the instruments used to assess student competencies unfairly underestimate their ability levels.

The Test Design Project studies in test bias were initiated by the Center for the Study of Evaluation with the belief that "bias" in assessment occurs in both the nature of the test and the situation within which the test is given. These studies have used four primary approaches to identify and interpret such bias. The first asked whether the use of translated tests constitutes a viable strategy for assessing non-English speaking pupils. A well-regarded test of childrens' academic competencies, the Comprehensive Test of Basic Skills (CTBS), is widely used in both its original English version as well as a recent Spanish-language translation; would the CTBS and the CTBS-Español prove as successful as is claimed in being free from bias? This is a volatile issues not only from the viewpoint of statistical analyses but also in terms of the number of separate impacts

such tests have on education. The second approach was to determine whether the current schemes for isolating test and item bias could be successfully applied to datasets which might not necessarily meet the various theoretical and practical strictures those techniques require. Moreover, a substantial line of inquiry not intrinsically statistical in nature, i.e., content analysis and linguistic analysis, was seen as central to the task of isolating both the fact and to some extent the root causes of bias. The third primary focus of effort in the analysis of test bias was to see whether new methods, potentially more direct and more amenable to use in the field, could be as successful in detection and interpretation of bias as previous state-of-the-art analytic schemes which have been suggested. One strong objection to many of those schemes has been that few of them yield unambiguous information about bias, and most are substantially more complex in their execution than might be desirable. A fourth approach, related to the others, examined the likelihood that test bias also occurs in alternative forms of testing, such as the free writing of prose in response to a prompt. Here the potential for bias extends not only to the questions of individual examinee performance but also to the performance of the persons who rate the essays afterwards.

Following a year's planning and data acquisition, the second year of the study was devoted to four separate analyses addressing the goals noted above:

(1) Conventional analyses of bias including the approaches suggested by classical test theory and newer multidimentional scaling methods, coupled with pursuit of simplifications to the statistical analysis of bias,

including, first, the model of partitioned variances and, second, the Student-Problem (S-P) methods in use in Japan (Sato and Kurata, 1977; Sata, 1980);

2) Analysis of selected aspects of item content, the extent of linguistic, cultural and social bias imbedded within item stems and answers, and the quality of translation of the CTBS from its original English version to the Spanish-language CTBS-Español;

3) Analysis of a selected dataset within the test bias project which contained scores of both the English- and Spanish-language versions of the CTBS from the same set of students, judged by their teachers to be equally competent in both English and Spanish; and

4) Analysis of ratings made by Hispanic and non-Hispanic raters using an objective scoring system, who reviewed a special set of essays generated by groups of Hispanic and non-Hispanic primary school students.

Certain aspects of analysis suggested in the original proposal for this project contained in some detail in interim reports were initially very appealing as possible routes for optimizing the detection and interpretation of bias.

They are the partitioning of variances, log-linear analysis, and multi-dimensional scaling even when the necessary initial specifications are not known with sufficient accuracy. However, in the long run each of these proved to be theoretically problematic, relatively cumbersome and statistically unwise in their application to the test bias question.

Partitioning of variance into a between-class and a within-class component formed one major aspect of the original effort, and was presented in some detail in the November 1980 deliverable. The primary intent was to utilize such partitioning for each item to reveal patterns of bias

through the interpretation of both relative sizes of the variances and their correlations with total test score and popular distractor answers. However, substantive theoretical and practical arguments became evident as work progressed. The first is that the examination of within-class variations within items runs a large risk of violating assumptions of homoscedasticity, and such violations cannot be resolved independently. Secondly, between-class variations are actually non-orthogonal to within-class variations except in certain situations. A desirable index of bias would be one which utilizes information from that portion of between-class variation which excludes all other portions of the variance, but this index proved to be intractable in practice. Additionally, calculations of effect size in relation to variations in class size contain some unsolvable unknowns. Log-linear analysis has been successfully applied to a number of studies in sociology and was considered as a viable tool for analysis of bias until it was determined that the stability of computations involved in this technique was questionable with the sample sizes of the data sets available. Additionally, the interpretation of results in the context either of specific items or specific examinees (and thus a logical route to the isolating of item bias) was hampered by requirements for secondary analyses following the initial solution. The presence of inadequate sample size profoundly affects the utility of multidimensional scaling in these investigations. Likewise, the issue of computational indeterminacy was a noticable hinderance to effective solutions using that method. When certain rigorous conditions for specification of initial parameters are met, multidimensional scaling may well prove as effective in detecting bias as other techniques.

The S-P method, first discussed in the November 1980 deliverable, appears to hold a number of possibilities for effective and unambiguous analysis of bias. It is a highly versatile contribution to the field of testing from Japan, and contains minimal requirements on sample size, prior scoring, item scaling and the like. The S-P model lends itself to extensions into nondichotomous scoring and multiple pattern analysis, as well as the possibility that the role of guessing in achievement test scores can be analyzed effectively. In the main, the most recent efforts of this project have been directed at isolating sources of patterning in examinees' responses as a function of test items and distractors, student abilities and backgrounds, and their interactions. Towards this end, the S-P method, coupled with a small number of other techniques, has proved singularly successful at the task of determining degree of item bias, and the method of content analysis has proved to be an important contribution to understanding the language used in the test.

The analyses conducted during the year have been prepared for publication as follows:

McArthur, D.L. Detection of item bias using analyses of response patterns, Summer, 1981. Submitted for publication to the Journal of Educational Measurement (Appendix A).

McArthur, D.L. Performance patterns of bilingual children tested in both languages, Summer, 1981. Submitted for publication to the Journal of Educational Measurement (Appendix B).

McArthur, D.L. Bias in the writing of prose and its appraisal, Fall, 1981. To be submitted (Appendix C).

Cabello, B. Potential sources of bias in dual language achievement tests, Fall, 1981. To be submitted to TESOL Quarterly (Appendix D).

Cabello, B Cultural interference in reading comprehension: An alternative explanation, Fall, 1981. Accepted for the Annual Meeting of the California Educational Research Association, San Diego, November, 1981 (Appendix E).

Portions of these papers also have been accepted for the Annual Meeting of the American Educational Research Association, New York, Spring, 1982.

DETECTION OF ITEM BIAS USING ANALYSES OF RESPONSE PATTERNS

David L. McArthur
Center for the Study of Evaluation
University of California Los Angeles

Supported by a grant from the National Institute of Education (NIE-G-80-0012). Appreciation is extended to Beverly Cabello for her analysis of cultural and linguistic issues.

Abstract

Item bias, when present in a multiple-choice test, can be detected by appropriate analyses of the persons x items scoring matrix. Five related schemes for the statistical analysis of bias were applied to a widely used primary skills multiple-choice test which was administered in either its English or Spanish-language version at each of the two levels, to 1250 students in bilingual education programs. The results indicate that from one-fifth to one-third of the items in the tests show strong evidence of bias, corroborated by a separate analysis of linguistic and cultural sources of bias for both the biased items and those items with no statistical findings of bias.

A systematic but unanticipated pattern of responses to a multiple-choice test found for an entire group of test-takers is generally regarded as evidence of bias. This interpretation results from indications of one or more differences between groups on levels of knowledge and skill, or in linguistic and cultural issues related to the use of language in the test. However, the behaviors of individual respondents have important consequences for that interpretation. Whether the respondent unerringly picks the correct response, or successfully engages in elimination of incorrect answers, or guesses well, the observer scores the item "correct" and concludes that the student "knows" the required skills or material. The inference that the respondent "does not know" is made whether he/she guesses incorrectly, eliminates wrong choices badly, or chooses an attractive but incorrect alternative.

Most likely, what look like systematic patterns of bias in test items are the results of complex interactions of these group and individual factors with one another and with certain properties of the test items. What is required to make sense of the issue of bias is analysis of patterns found in these combinations of performance. The multiplicity of possible patterns suggests that the detection and interpretation of bias must be conducted along several routes.

Goals of this research

The first of two purposes of this paper is to investigate analyses of the persons x items scoring matrix of a test for the detection of item bias. The persons x items scoring matrix contains a significant amount of information about the patterns of responses generated by a set of examinees.

Using a few geometrical and statistical considerations, the patterns of responses from separate groups of examinees tested with the same instrument can be compared. If these patterns show that the test is not measuring the same thing--skills, competence, thinking abilities--in comparable groups, if the groups are responding to different aspects of the test items, or if cultural and/or linguistic issues take precedence, it may be that the test is biased.

The second purpose of this paper is to study empirically the question of bias as shown by these several techniques in the context of a widely used achievement test, the Comprehensive Test of Basic Skills (CTBS), which has been translated from English into Spanish. The claims made about this instrument include that statement that the Spanish-language version represents a close replicate of the English-language version with careful attention having been exercised in removing all forms of unintended bias. The primary task of this analysis is to ascertain the degree of comparability of the two versions of the CTBS in the assessment of similar groups of children, and to see if any bias remains.

Related literature

A substantial research literature has developed around the term "item bias" in the search for a single best all-purpose indicator which always reveals bias whenever systematic discrepancies in performance between groups are found. A large number of methods have been proposed and a large number of studies conducted (cf. reviews in Berk, in press; Subkoviak, Mack and Ironson, 1981). Certain tests such as the Wechsler Intelligence Scale for Children have been extensively investigated

(cf. Sandoval, 1979). The range of applications of the term "bias" is quite broad: studies have examined sociocultural bias and the stereotyping of items and answers, cultural differences, and linguistic variations (cf. Jensen, 1980); construct bias and the different aspects of performance tapped in different examinee groups by the same test (cf. Ebel, 1975); and contextual bias and the misuse of tests with specific groups (cf. Williams, 1971). Occasionally the word is even used to mean a conscious preference on the part of the examinee (Hudson, 1963).

Increasingly complex techniques have been set forth for the detection of bias in items. Methods have been based on analysis of variance, transformed item difficulties, factor techniques, adjusted chi square procedures, distractor analyses, "adverse impact" and item characteristic curves (Merz, 1980; Petersen, 1980; Rudner, Getson and Knight, 1980). Many of these methods are statistically complex but, with the exception of the last, statistically inelegant (Hunter, 1975); unfortunately the most elegant solution, item characteristic curve analysis, requires large numbers of items and respondents for its computation. Few of these approaches offer convincing or useful explanations of why some items are biased and others are not (Crowder, 1979). Faced with the multiplicity of both the forms of item bias and the statistical methods that have been put forward to detect such bias, one logical place to begin is to inquire about the nature of a test which is absolutely free of bias.

An unbiased test

If a test could be created which fulfilled all of the requirements of a bias-free instrument, its items would all measure the same trait or

ability and be equally reliable and equally valid for all groups (Petersen, 1980). It would also show orderly variation in the relative difficulties of the items, and be responded to in an orderly manner by every individual. One example of the outcome of this improbable creature is the familiar perfect Guttman scale, in which persons are perfectly ordered by increments of skill level, and items within the test are perfectly ordered by increments of difficulty. No higher-level item is mastered by any respondent until each lower-level item is mastered; guessing also plays no role. The sequence of successes and failures is highly deterministic.

Figure 1A represents a ten-item test with right/wrong scores for ten respondents. These ten persons never successfully answered a more

---

Insert Figures 1A and 1B about here

---

difficult item without first having succeeded on a less difficult item. An axis of performance can be drawn on the diagonal to separate all correct scores from all incorrect scores. While the total p-value for the test is lower for another group of ten persons tested on the same ten items, shown in Figure 1 B, the performance patterns are parallel. Other than a main effect due to groups, nowhere in either diagram is any indication of a systematic unexpected difference in the pattern of responses or bias in the test.

A slightly biased test

A somewhat less artificial example of test results from a multiple-choice test is shown in figure 2A; the score matrix of a hypothetical

---

Insert Figures 2A and 2B about here

---

ten-item test has been sorted by both persons, on ascending total score, and by items, on ascending level of difficulty. Neither persons nor items is perfectly ordered in the sense used above, and guessing of correct answers probably contributes by an unknown amount to the scores obtained. Not one but two dividing lines are now required to separate the patterns of performance in this figure. The first line, a cumulative ogive representing student performance, is drawn on the matrix based on the total correct score for every respondent. The second, representing problem difficulty, is drawn as a cumulative ogive based on item p-values. Note that for a test which demonstrates exclusively random responding, the theoretical position of the student curve (S-curve) would be vertical, and of the problem curve (P-curve), horizontal.

At this juncture we introduce a second set of data obtained from the same hypothetical test. The "respondents" were slightly less capable on most items but all other considerations were held equal. A score matrix for the same set of items as shown in Figure 2A but the second group of examinees is shown in Figure 2B. The relative order of items is somewhat changed because of differing levels of difficulty; the second group performs less well overall than the first group. Statistical differences between the data in Figures 2A and 2B should reflect overall item and group differences, but because of the idealized symmetry between the two, there is little likelihood that a statistical indicator of bias would prove significant. An initial analysis of these figures recommended by Jensen (1980) is a two factor (group x items) nested analysis of variance. The interpretation of a significant groups effect, in the absence of

other significant factors, is that the groups behave symmetrically with respect to ordering of item difficulties but that one group is consistently more capable across the trait being appraised by this test. A significant difference on both the groups and items factors, plus a significant interaction between groups and items, together suggest that the test items and examinee abilities in the two groups are heterogeneous.[1] However, these findings would be quite insufficient to say that the test is biased (Hunter, 1975), and, additionally, do not account for the contribution of guessing.

A second approach recommended by Jensen (1980) for understanding the differences between the two figures uses the phi coefficient, which is the correlation obtained between the group response to a given item and the same group's response to any other item in the test. Phi is a measure of joint contingency; Jensen explains its use for analysis of bias:

> Only if the two items have the same difficulty...can phi be equal to 1....To determine the intrinsic correlation (of the items) free of the influences in item difficulty, we must divide the obtained phi by the maximum value of phi that could possibly be obtained with the given marginal frequencies(p.431).

The ratio of phi to maximum value of phi is summed over all possible pairs of items for each group, and then the ratios are compared. The null hypothesis for this comparison is that the difference between the obtained sums is not different from randomness, and thus there is no systematic discrepancy in group performance. In the artificial situation shown by the Guttman scale for both groups in Figure 1, this test is necessarily nonsignificant. For data which does not fit the mandates of a perfect

scale, the obtained value for the comparison of ratio sums increases as the discrepancy in overall patterns of response by the two separate groups widens.[2] While the amount of difference between groups is given by the analyses of variance and phi, the nature of patterns of response to items is not adequately explained.

Only a small number of statistically-based analyses specifically designed to study patterns of responding to multiple-choice tests have been proposed. Tatsuoka (1981) and Harnisch and Linn (1981) have been working on a norm conformity index and other parameters which address each individual's performance in the context of patterns obtained by all members of the group. Sato (1980) defines an index of disparity between actual and ideal response patterns which can be applied to individuals or to items. To unravel the problem of patterns, we now turn to Sato's system of analysis of the persons x items matrix.

The S-P method and analysis of the persons x items matrix

The key element in Sato's (1980) S-P method of analysis of test performance is the doubly-ordered persons x items matrix, with student curve (S-curve) and problem curve (P-curve) drawn in. In Japan, this procedure is widely used in classrooms to obtain the characteristic performance of the set of examinees, which may be compared visually to several "standard" curve functions for diagnostic purposes.[3]

Sato has developed an index of discrepancy to evaluate the degree to which the S and P curves do not conform either to one another or to the Guttman scale. Except in the case of the perfectly ordered sets

shown in Figure 1, there is always some degree of discrepancy between curves. The index is explained as follows:

$$D^* = \frac{A\ (\ N,\ n,\ \bar{p})}{A_B(\ N,\ n,\ \bar{p})}$$ where the denominator

> ...is the area between the S curve and the P curve in the given S-P chart for a group of N students who took n-problem test and got an average problem-passing rate $\bar{p}$, and $A_B(\ N,\ n,\ \bar{p}\ )$ is the area between the two curves as modeled by cumulative binomial distributions with parameters N, n, and $\bar{p}$, respectively (Sato, 1980, p. 15).

The denominator is a function which expresses a truly random pattern of responses for a test with a given number of subjects, given number of items, and given average passing rate, while the numerator reflects the obtained pattern for that test. As the value of this ratio approaches 1.0, it portrays an increasingly random pattern of responses. For the perfect Guttman scale as represented by Figure 1, the numerator will be 0 and thus D* will be 0.[4]

Indices of discrepancy, when computed for each of two groups of examinees, may not be statistically compared because of differences in ranking of item difficulty, and/or compound differences in response patterns to several items. However, as long as the two D* values obtained are not equivalent, it is an indication that somewhere within the matrices are one or more items which are behaving dissimilarly across groups.

Analysis of respondents above P curve

Patterns of discrepant performance result from a mixture of random behaviors and wrong choices, except for those items which are so easy that no respondent gets them wrong. Aside from the tautology that respondents with less ability are less likely to answer a given item correctly,

all other things being equal they are also likely to use chance responding. Analysis of those respondents who are unlikely to be answering randomly would seem a likely means to understanding patterns and bias in items. To begin constructing a simple analytic solution to this problem, suppose we take a single uncomplicated item from the S-P chart, and examine the pattern of responses for only that portion of the same group of examinees for whom the prediction of success is relatively high, i.e., those above the P-curve. These are the examinees who tended to score better overall. Specifically, respondents at the very top of this select subgroup are expected to have had a finite but small probability of having guessed their way to success. Respondents at the bottom of this select subgroup would have a finitely larger probability, while those at the very bottom of the entire S-P chart would be likely to have a more random pattern.

If the selected item, however, is one for which no individual within the sample, no matter how skilled, is able to answer knowledgeably, the response pattern among the select group of putative "masters" should be random, and should not differ from the response pattern of those examinees not included in this subgroup. For a four-choice item of this kind, the item's p-value should be about .25, and the select subgroup of putative "masters" would be correct only 25% of the time. Figure 3 illustrates a pattern of responses for a nearly random item, in contrast with an item which is fairly well-fitted to the skills of a set of respondents.

Insert Figure 3 about here

The proportions of "masters" who are indeed correct can be compared between groups. With relatively uniform variances, the test of significant difference in independent proportions applied to this problem yields a z

score; a significant z score would be an indication of possible bias separate from the difference in average passing rates for that item, if any. A comparison of nonuniform variances requires transforming the item difficulties into standard score form, then testing the size of the difference following Rudner, Getson and Knight, (1980). Within certain limits, an item which is relatively easy for one group and relatively difficult for another may show no bias in the proportions of "masters" who are correct, because those individuals who place above the P curve all have the ability to answer that item correctly. However, on another item one of two groups may not be academically equipped, or may be prevented by responding by biases in the test, curriculum or culture; thus the proportions may differ, possibly by an amount sufficiently large to be deemed significant.

Analysis of distractors

One further analysis of the potentially biased item is to examine the patterns of wrong answers made by the separate groups of respondents. Within the multiple choice test format, differences between groups in the attractiveness of incorrect responses signal that the item's wrong choices may be differentially distracting. When a given item has attractive but incorrect responses for one group, Goodman and Kruskal's lambda indicates whether another group shares the same proportional pattern of selecting those incorrect responses (Veale and Forman, 1976). Lambda is an index of predictive association, which shows "...how one is led to predict differentially in light of the relationship..."(Hayes, 1963, p. 610, italics original). It is calculated for a problem in-

volving two groups by evaluating the largest discrepancy between rates of responding to similar wrong choices:

$$\lambda = \frac{\Sigma \text{ max.}f_{jk} - \text{max.}f_{.k}}{N - \text{max-}f_{.k}}$$

where $\text{max.}f_{jk}$ is the larger frequency of the two groups for any single wrong choice, and $\text{max.}f_{.k}$ is the larger marginal frequency of the two groups summed across all wrong choices.

If Goodman and Kruskal's lambda is appreciably above zero, the interpretation can be made that the pattern of distraction is different for the two groups. If the index is zero, even though the difficulty of the item and/or the proportions who select a wrong option may differ between the two groups, the pattern of selecting the wrong answers is about the same.

Another check on the relative attractiveness of a wrong answer can be made by counting the number of wrong answers which are chosen at least 10% more often than the next most popular wrong answers. These particular wrong choices constitute a class of "popular distractors," each of which can be studied further. The easiest comparison is between those items for which both groups picked the same popular distractor and those items for which both groups picked different popular distractors. Note that in this latter case, the computation of lambda will always yield a nonzero value.

A series of analyses of item bias has been described, with special attention paid to those comparisons premised on the persons x items scoring matrix, doubly sorted. The following sections describe the execution of these analyses in the context of a multi-language achievement test.

Method

Instruments

For a study of the possible bias inherent in a multi-language test, two levels of the Comprehensive Test of Basic Skills (CTBS) published by CTB/McGraw Hill (1974, 1978) were administered in this study. Students in grades 2 and 3 were given the CTBS Level C; participating fifth and sixth grade students took Level 2. CTBS-English Level C is designed for students in grades 1.6-2.9; CTBS-Spanish Level C is designed for students in grade 2. CTBS-English level 2 has a target population in grades 4.5 to 6.9; the Spanish translation was designed for students in grades 5 and 6.

The CTBS-English and CTBS-Spanish tests were selected for several reasons. Test content is roughly parallel. The CTBS-Spanish was the first test at CTB/McGraw Hill to be subjected to a four-step editorial procedure designed to reduce test bias; included were studies of content validity, application of editorial guidelines in item construction, reviews for bias, and separate ethnic group pilot studies with the test. In the translation of the CTBS from English to Spanish, the test developers tried to keep the test content and measurement features intact. This, of course, meant that in some cases word-for-word translations were not possible. Nevertheless, it was the intent of the publisher to provide tests that are similar in rationale and in the process/content classification scheme. Thus, both the English- and Spanish-language versions used in this study purport to measure the following objectives:

1. the ability to recognize or recall information
2. the ability to translate or convert concepts from one kind of language (verbal or symbolic) to another

3. the ability to comprehend concepts and their interrelationships
4. the ability to apply techniques, including performing operations.
5. the ability to extend interpretation beyond stated information (CTBS, 1974/1978)

Test length, test time and adminstration procedures are exactly the same for English and Spanish versions of each test level.

Subjects

Five school districts in the state of California participated in the study. The total number of pupils tested was 1259, representing 81 intact classrooms.

Classrooms were selected to represent a wide range of program options. The criterion for selection of school districts was that they had bilingual-bicultural education programs funded either by Title VIII or by the ESEA. Potential participants were identified from schools listed in the California State Department of Education 1979 Bilingual Program Directory. From this list, invitations were sent to schools which had at least two classes at the same grade level (grades one, two, five, or six) having bilingual programs. Additionally, instruction had to be delivered in self-contained, multi-subject settings; departmentalized or pull-out programs were excluded.

Analyses

Five statistics explained above were used to evaluate the data for every item separately. Each uses a minimum threshhold value, above which the result is taken as an indication of possible bias in the item. The analyses and their minimums can be summarized as follows:

a) Test of proportions of correct scores: across groups, a difference between transformed p-values which generates a $z>1.96$;

b) Test of proportions of correct scores for "masters": across groups, a difference between proportions of those respondents above the P-curve who make errors, which generates a $z>1.96$;

c) Test of chance responding by "masters": within each group, a difference between the obtained proportion of those passing the item and a theoretical p-value of .25, which generates a $z<1.96$;

d) Test of differential attractiveness of wrong answers: a Goodman and Kruskal's lambda computed on the proportions of incorrect answers by choice within item, such that $\lambda>0.0$;

e) Test of popular distractors: a wrong choice for an item attracting at least 10% or more responses than the next most popular wrong choice for that item.

Results

The number of items within each subtest by level, and the number of students in each of two language groups who were included are shown at the top of Table 1. Item p-values indicate that items ranged from moderately

---

Insert Table 1 about here

---

easy to very difficult for both language groups, with a overall mean of somewhat over half of the items correct. While in a few items the Spanish-language group did better, without exception the Spanish-language groups always scored lower overall on the subtests. In every instance the maximum p-values achieved by the English-language groups are slightly higher than the comparable scores for the Spanish lanugage groups. Table 1 also shows for the corresponding number of students, the p-value needed for a significant ($p<.05$) difference from chance responding to an item. This figure is obtained by reversing the usual computation for the test of independent proportions, using $z = 1.96$ and $p_{chance} = .25$. For all but one of the subtests, both language groups had one or more items which appear

to represent random choice of the correct answer. Except for the Passage Comprehension subtest at Level C, the Spanish-language group appears to make random selections more often than the English-language group, an assumption which is explored further below.

For purposes of illustration, two analyses recommended by Jensen (1980) were conducted on the subtest with the smallest number of items, Level C Passage Comprehension. The two-factor nested analysis of variance for this subtest shows a significant effect due to the groups factor ($F(1,650) = 54.91$, MSerror = 1.37), and a significant effect due to the interaction between items and groups ($F(17,11050) = 2.61$, MSerror = 0.43). The ratio of phi to phi-max is higher for the English-language sample than for the Spanish-language sample (English mean ϕ/ϕ-max = .8207; Spanish mean ϕ/ϕ-max = .7666, $t(151) = 4.01$, $p<.01$). This brief set of findings indicates only that the language groups are not performing the same way as one another on the subtest. It seems that the Spanish-language sample may have had more difficulty with some items than did their English-language counterparts. No further detail can be learned from these analyses, and they are not used in the study of the remaining subtests.

The S-P charts were drafted for each subtest by language group for a total of eight complete charts. The index of discrepancy D* is presented in the last row of Table 1. The fact that the D* values are higher for the Spanish-language groups suggests that they engaged in patterns closer to chance responding more often than did English-language groups. While the differences between pairs of D* values are large for the Passage Comprehension subtest at both level C and level 2, these values cannot be compared further.

The specific reasons why the Spanish-language versions generate larger D* values can only be made evident with further analyses.

Results from the set of five analyses which together provide sufficient evidence of patterns of discrepant performance are presented below and in Table 2. The table shows percentages of items for each of the four subtests in this study which exceed a critical minimum on each of the five analyses.

Test of proportions of correct scores. The first of the concise set of analyses is the test of proportions, which is applicable to percentages of correct answers expressed in standard score form, for both groups on each item of each subtest. The first two rows of Table 2 show the percent of items favoring the English- or Spanish-language groups. Six out of every ten items in the Vocabulary subtests show significant

---

Insert Table 2 about here

---

differences between groups; in a majority of instances the higher group is always the English-language group. Half of the items in the Passage Comprehension subtest at Level C show a significant difference and over three-quarters of the items in that subtest at Level 2 show a significant difference; in no instance are the Spanish-language groups ahead of their English-language counterparts.

Test of proportions of correct scores for "masters." Both the second and third analyses in this set are based on the selective sample of "masters," those students whose overall scoring position places them above the P-curve for each item. By evaluating the proportions of correct scores for those members of the language groups, a list of statistically significant discrepancies between "masters" is generated. The third and

fourth rows of Table 2 show the percent of items within subtest for which the success rate among "masters" is significantly higher for the English-language or Spanish-language groups. The Passage Comprehension subtests at both levels appear to have different rates at which the "masters" are able to avoid the wrong answer; in the majority of instances the rate is higher for the English-language groups. In the Passage Comprehension subtests, the rate is uniformly higher for the English-language groups.

Test of chance responding by "masters." How often the samples of "masters" are not able to choose the correct response at a rate better than chance forms a third part of the analysis. The fifth and sixth rows of Table 2 show that for the Level C subtests, no items are found for which either group responded randomly. However, for Level 2, a small number of items in both subtests elicited chance responding by "masters". These items appear to be so difficult that not even the better students could knowledgeably select the correct response. The Spanish-language group has a much larger number of chance responses among "masters" the English-language groups on the Level 2 Passage Comprehension subtest.

Test of differential attractiveness of wrong answers. The fourth analysis in this sequence is the analysis of differential patterns of incorrect responses. Goodman and Kruskal's lambda was calculated for each item, using a 2 x 3 table of groups by incorrect response rates. Values ranged from 0.0 to .23, with a median of 0. Lambda will be 0 for any 2 x 3 table of proportions for which both groups are attracted to the same response, even if the actual dimensions of those attractions differ drastically. As there is no exact test of significance, any non-

zero lambda was considered to be an indicator of possible bias. The seventh row of Table 2 shows the percentage of items within each subtest for which a nonzero lambda was found. The ratio of such items to the number of items within subtest ranges from 1:4 to 1:2, suggesting that, when wrong answers were selected the two languages groups often behaved very differently.

Test of popular distractors. The concluding analysis in this series asks whether there are any incorrect choices which were sufficiently attractive to be classed as popular distractors. In the final rows of Table 2 are shown the percentage of items which meet the 10%-or-greater criterion for the English-language groups, the Spanish-language groups, and jointly across groups. Except in Passage Comprehension at level 2, the Spanish-language group's results show more items with popular distractors than the English-language group. Percent joint overlap is of particular interest, since that value gives another indication of the uniformity of behaviors across language groups when selecting incorrect responses. In the subtests in this study, the joint overlap of popular distractors is very small, suggesting again that many items of the English version of the test and the Spanish translation may not be as comparable as the test designers intended.

The degree of overlap between the five analyses in terms of the number of positive findings for each subtest is shown in Table 3. The

---

Insert Table 3 about here

---

percentage of items for which none of the preceding analyses show evidence of bias is remarkably small. Level C Passage Comprehension, for example, has only a single item which never shows a difference between the language

groups. Over half of the items in that subtest have at least two positive findings, and four of the items have three positive findings. Table 3 shows that the percentage of items for which three, four, or five out of five statistical indicators yield positive results varies from about one-fifth to about two-fifths of the items within each subtest.

Content analysis

On the basis of the preceding evidence from the statistical approach to bias detection in the CTBS, those items which show agreement of three or more indicators were subjected to a careful analysis of item content. The content analysis was a search for possible linguistic, curricular, and/or cultural reasons which might explain differential performance between language groups. This portion of the study was undertaken by an educational researcher fluent in both English and Spanish, making extensive reference to the curricular materials used by the students in the sample, and consulting with native speakers of various dialects in making an appraisal. Five categories were tabulated as possible sources of influence which item content might exert on the different language groups:

a) Mistranslation: the meaning and/or grammatical form of a key word or phrase within the item was translated from the English original in a manner which is an incorrect or inappropriate use of the Spanish language;

b) Cultural bias: some key word or phrase within the item requires familiarity with objects, behaviors, or values which are not normally found in the Spanish and Latino cultures, or which may have very different interpretations;

c) Linguistic bias: some key word or phrase within the item requires familiarity with an idiomatic expression or verbal allusions which, because of innate differences in language, do not translate well;

d) Low frequency word bias: some key word or phrase within the item is not found, or rarely found, in the basal readers used for instruction by the students in our sample.

e) Unfamiliar context bias: some key word or phrase within the item appears in a context which is quite different from that found for the word or phrase in the basal readers used for instruction.

An example of item content judged to bias respondents is shown by item number 29 of the Level C Vocabulary subtest, an item for which all statistical indicators point to possible trouble. Item 29 (rated as category c, linguistic bias) requires the student to select a synonym for "happy." The English-language version of the test yielded responses which appear significantly disadvantaged on this particular item. While the correct option for this item in the Spanish-language version, /alegre/, was selected 60% of the time by our sample, the correct option in the English-language version, /gay/, was selected only by 13% of the sample. The English-language respondents instead split their selection equally between two of the remaining options. Only one other item in the entire test set received as strong a rejection, suggesting that among second and third graders, the slang English-language meaning for 'gay' has not only rendered it useless as a synonym for 'happy' but has given it a strong pejorative flavor as well.

Table 4 shows data for items in each of the four subtests for which

---

Insert Table 4 about here

---

the content analysis identified probable sources of bias. The entries in the table represent tabulations of the content analysis categories for those items on each subtest which have three or more statistical indicators. For the Level C Vocabulary subtest, twelve items have at least three statistical indicators; nine of those twelve show evidence of linguistic

bias, and five of the nine show evidence from an additional category of content bias as well. Three of the four items from the Level C Passage Comprehension subtest fit at least one of the categories of content bias, two of them with multiple indicators. Only four out of nineteen on the Level 2 Vocabulary subtest items with three or more statistical indicators do not have ostensible problems as shown by the content analysis procedure. Of twenty-one items in the Level 2 Passage Comprehension subtest with three or more indicators, only three cannot be corroborated by the analysis of content. None of the items in any subtest which had no statistical indicators of bias were found to have any content indicators of bias.

Table 5 presents a summary of subtest performance by group when those items for which three or more statistical indicators turn up positive

---

Insert Table 5 about here

---

are excluded. In three of the four subtests, the adjusted scores of the Spanish-language groups move closer to their English-language counterparts. A substantial difference remains, however, between scores for the Passage Comprehension subtest at Level 2. The gain from initial to adjusted group mean by the Spanish-language group is quite insufficient to raise that value to the level of the English-language group. The adjusted minimum p-values achieved by both groups move upward but the English-language group pulls ahead noticeably.

Discussion

Five relatively simple analyses have been presented which point to five related considerations in the search for bias. These are a) overall group differences and their direction, b) differences in performance by a

select subsample of better respondents within groups, c) differences from chance responding by those subsamples, d) differences between groups in the selection of wrong answers, and e) degree of distraction provided by wrong item choices. The first of these follows the well-known Anghoff delta procedure (Anghoff, 1972), without resorting to the arbitrary use of rescaling, which simply serves for added convenience. The second and third analyses make use of the select subsample of putative "masters", those students within each group whose overall performances place them above the P-curve; these approaches are extensions of the work of Sato (1980) and colleagues. The fourth and fifth procedures examine the bias question by studying those parts of the multiple-choice item which are usually excluded from study in a right-wrong scoring context (cf. Powell and Isbister, 1974).

For purposes of this paper, the five procedures are considered jointly, with equal weights. Interpretations of bias are confirmed in the clear majority of cases where the joint indication of three or more statistics is found for an item. Certain problems remain to be solved, however, and therefore some conditions must be placed on the use of this set of approaches to the detection of item bias. It is clear, for example, that the first index, because it is based on proportion of correct items, is to be used with caution: "proportions of correct answers in a group of examinees is not really a measure of item difficulty. This proportion describes not only the test item but also the group tested" (Lord, 1980, p. 35). Indeed, throughout it must be remembered that the results of this study are descriptive of this sample only, and no external criteria are available to evaluate comparability across language groups by grade.

A second objection is that the psychometric properties of the CTBS items are only partially expressed by reliance on p-values and the S-P chart, which at its core relies on the index of item difficulty. Thus, the conclusions drawn from work with that chart are only as good as the strength of the item difficulty metric. In addition, the S-P chart suffers from other metric problems. The first is that the doubly-sorted persons x items matrix treats data, in part, as interval rather than continuous data. Thus, for instance, subtle gradations of difficulty may be given the same credance as larger differences in the case where p-values are nonuniformly distributed. Analogously, nonlinear distributions of total performance scores may contribute in unknown ways to the use made of ranking information regarding respondents: the patterns may not be as smooth as the chart makes them appear. Moreover, as the S-P chart approaches randomness and its index of discrepancy, D*, approaches 1.0, increasingly complex but hidden interactions between the properties of the items in the test and the attributes of the sample are likely. Thus, the second and third statistics in the analytic set depend upon certain assumptions about the nature of performance pattens, violations of which bear rather unclear consequences. Related problems appear in item characteristic curve analysis (Linn, Levine, Hasting, and Wardrop, 1980), and in the "adverse impact" approach (Merz, 1980).

A third objection to the procedures used in this study centers on issues of guessing. In the absence of an externally valid explicit criterion, correction for guessing does not seem feasible (Choppin, 1974). Yet assumptions about the occurence and distribution of guessing affect all aspects of the analysis, particularly statistics which address incorrect

responses. Volitional bias, quite likely contributing to the anomalous response by the English-language group to item 29 on the Level C Vocabulary subtest, is nowhere adequately considered. How much of a role guessing plays is not well treated by the assumption that chance responding is represented by p = .25. In the very likely event that some members of any group will engage in guessing some of the time on some items, only the most general and simplistic conclusions can be drawn from the data presented here. One problem of particular note is the strong possibility that guessing assumes a gradiant distribution within the person x items matrix. That is, from the most capable to the least capable person, the contribution of guessing on any item may move from relatively low probability to relatively high probability, thus potentially interfering with diagnosis of problems inherent in the item. But such diagnosis lies at the heart of the effort to decipher and describe item bias. Until the gradiant problem is separated from the bias problem, only partially satisfactory conclusions can be drawn about either.

On the positive side, the high level of match between content analysis and the aggregate of statistical evidence suggests that this simple approach to bias detection may have as much viability as more laborious and unwieldy procedures. The ease of computations and interpretations, and the parsimony of explanation are also favorable points (Merz, 1980). While some attempt is made in the preceding pages to demonstrate the use of multiple indicators, more possibilities can be pursued within this framework. The explanatory power of the five-part procedure appears to exceed that offered by analysis of variance or phi/phi-max, and the assumptions

required about the configuration of persons and items are fewer in number than those required by the modified chi-square analyses which recently have been challenged as inadequate (Marascuilo and Slaughter, in press).

Comparison of the present set of results with those of more complex analytic procedures conducted on the same data set awaits further study. However, unlike the results reported by Linn, Levine, Hastings and Wardrop (1981), in which item characteristic curve analyses for a hypothetical dataset "...did not lend themselves to making generalizations about features of items...(p. 38)," the findings of the present study suggest at least one concluding observation. Many signals point to a primary conclusion that a number of items in the English-language and Spanish-language versions of the CTBS do not seem to be comparable. Across a spectrum of indicators, the Spanish-language groups regularly produced lower scores. In three of four subtests, removing those items for which three or more statistical indicators pointed to difficulty gave adjusted scores which were very similar between groups. In the fourth subtest, that correction did not yield significant improvement, suggesting that the Spanish-language sample at grade 6 may be disadvantaged in some respect unrelated to the CTBS itself.

Footnotes

[1]The comparison of Figures 2A and 2B yields only a significant difference on the factor of items ($F(9,162)=13.98$, $p<.001$).

[2]For the difference between Figures 2A and 2B, $\chi^2=8.0222$, $p<.01$.

[3]Direct interpretation of item scores, person scores, and the amount of discrepancy between the S and P curves is relatively easy to accomplish; the same holds for item analysis, individual performance analysis, and other summary statistics within a group. In Japan, this system has been automated using a microcomputer (Sato, Takeya, Kurata, Morimoto and Chimura, 1981).

[4]In Figure 2A, $D^* = .2534$; in Figure 2B, $D^* = .3747$.

Table 1

Summary of performance by subtest by group

Subtest	Level C				Level 2			
	Vocabulary		Passage Comprehension		Vocabulary		Passage Comprehension	
Group	English	Spanish	English	Spanish	English	Spanish	English	Spanish
n items	33		18		40		45	
N students responding	364	286	363	280	378	231	377	203
$\bar{p}$ value	.6570	.6212	.6254	.5924	.5599	.4302	.5225	.3832
s.d.	.1619	.1775	.0874	.1139	.1473	.1506	.1254	.1022
maximum p	.8571	.8542	.7356	.7128	.8568	.7662	.7507	.6321
minimum p	.1395	.1538	.4826	.4088	.2892	.2078	.2366	.1272
minimum required p greater than chance responding	.2969	.3033	.2970	.3039	.2960	.3096	.2961	.3138
n items less than minimum required p	1	2	0	0	1	11	2	11
index of discrepancy D*	.3408	.3568	.2353	.4690	.4416	.4980	.4741	.6288

TABLE 2

Percentage of Items Exceeding
Critical Minimums in Five Analyses

Subtest	Level C		Level 2	
	Vocabulary	Passage Comprehension	Vocabulary	Passage Comprehension
Analysis				
a) Test of proportions of correct scores				
English significantly higher	45%	50%	55%	76%
Spanish significantly higher	18%	0%	8%	0%
b) Test of proportions of correct scores for "masters"				
English significantly higher	33%	44%	40%	60%
Spanish significantly higher	22%	0%	5%	0%
c) Test of chance responding by "masters"				
in English	0%	0%	3%	7%
in Spanish	0%	0%	3%	16%
d) Test of differential attractiveness of wrong answers between groups	36%	50%	43%	27%
e) Test of popular distractors				
in English	9%	11%	13%	29%
in Spanish	30%	17%	30%	24%
Overlap between groups	6%	0%	10%	13%

TABLE 3

Percent of Items Showing Statistical Indicators of Differential Performance

Subtest	Level C		Level 2	
	Vocabulary	Passage Comprehension	Vocabulary	Passage Comprehension
No indicators	9%	6%	23%	4%
One indicator	33%	39%	18%	20%
Two indicators	21%	33%	18%	40%
Three indicators	27%	22%	33%	34%
Four indicators	6%	0%	8%	2%
Five indicators	3%	0%	0%	0%

Table 4

Sources of content bias for items with three or more statistical indicators of differential performance, by subtest

Key:
a) test of proportions
b) test of proportions of correct scores for "masters"
c) test of chance responding by "masters"
d) test of differential attractiveness of wrong answers
e) test of popular distractors

1) mistranslation
2) cultural difference
3) linguistic difference
4) low frequency word or phrase
5) unfamiliar context for word or phrase

Level C Vocabulary	Level C Passage Comprehension	Level 2 Vocabulary	Level 2 Passage Comprehension
item 2 a b e; 4	item 1 a b d ; -	item 1 a b d ; 2	item 1 a b e; 1
6 a b d ; 3	4 a b e; 1 3	6 a b e; 1 3	2 a b e; -
7 a d e; 2 3	6 a d e; 2	8 a b d ; -	3 a b d e; 1 3
12 a b e; 3	7 a b d ; 2 3 4	9 a b e; 3 4	7 a b d ; 4 5
14 a b e; 4		11 a b e; 1 2	9 a c e; 4
15 a b c e; 1 3		12 a b d e; 1	15 a d e; 5
16 a b e; 3 5		13 a b e; 1 2 3	17 a b e; 2
20 b d e; 2 3		15 a b e; 2	18 a b c e; 4 5
23 a d e; 2 4		19 a b c e; 1 3	21 a b d ; 4 5
29 a b c d e; 3		20 a b d e; -	22 a b d ; 4 5
30 a b d e; 2 3		23 a d e; 2	24 a b c d ; -
32 a b e; 3		25 a b c d e; 1 2	25 a b c d ; 3
		26 a b c e; -	28 a b c e; 1
		32 c d e; 1	29 a c e; 1
		34 a c d ; -	34 a d e; 4 5
		35 a b c e; 4	36 a b c ; -
		36 b c e; 3	37 b c d ; 2
		39 a b c d ; 3	38 a b c e; 2
		40 a b d ; 3	39 a b c e; 4 5
			41 a b d ; 4 5
			45 a c d ; 3 4

Table 5

Revised summary of performance by subtest group, deleting items with three or more statistical indicators

Subtest	Level C				Level 2			
	Vocabulary		Passage Comprehension		Vocabulary		Passage Comprehension	
Group	English	Spanish	English	Spanish	English	Spanish	English	Spanish
adjusted n items	21		14		21		24	
adjusted mean	.6804	.6606	.6216	.6061	.5818	.5322	.5431	.4067
change from original	.0234	.0394	-.0038	.0137	.0219	.1020	.1230	.0969
adjusted s.d.	.1298	.1502	.0936	.1039	.1418	.1476	.0206	.0235
adjusted maximum	.8571	.8542	.7356	.7128	.8568	.7662	.7507	.5707
adjusted minimum	.4104	.3004	.4826	.4343	.3344	.3005	.2366	.1272

Figure Captions

Figures 1A and 1B: 1A) Perfect Guttman scale for a hypothetical ten-item test scored right (1) or wrong (0). Persons and items are uniformly ordered, by total correct score and level of difficulty, respectively. 1B) Perfect Guttman scale, showing uniform ordering with lower overall performance.

Figures 2A and 2B: 2A) Hypothetical score matrix for a ten-item test sorted by respondents on descending total score and by items on ascending level of difficulty. S- and P-curves reflect cumulative ogives of performance, and lead to an appraisal of the characteristic performance of the group. 1B) Hypothetical score matrix for the same test with a different group, again sorted by respondents and items.

Figure 3: Hypothetical patterns of response to two items by ten persons, showing a poorly-fitted and a better-fitted item.

1A)

Items	1	2	3	4	5	6	7	8	9	10	Total score
Persons A	1	1	1	1	1	1	1	1	1	1	10
B	1	1	1	1	1	1	1	1	1	0	9
C	1	1	1	1	1	1	1	1	0	0	8
D	1	1	1	1	1	1	1	0	0	0	7
E	1	1	1	1	1	1	0	0	0	0	6
F	1	1	1	1	1	0	0	0	0	0	5
G	1	1	1	1	0	0	0	0	0	0	4
H	1	1	1	0	0	0	0	0	0	0	3
I	1	1	0	0	0	0	0	0	0	0	2
J	1	0	0	0	0	0	0	0	0	0	1
% correct	100	90	80	70	60	50	40	30	20	10	

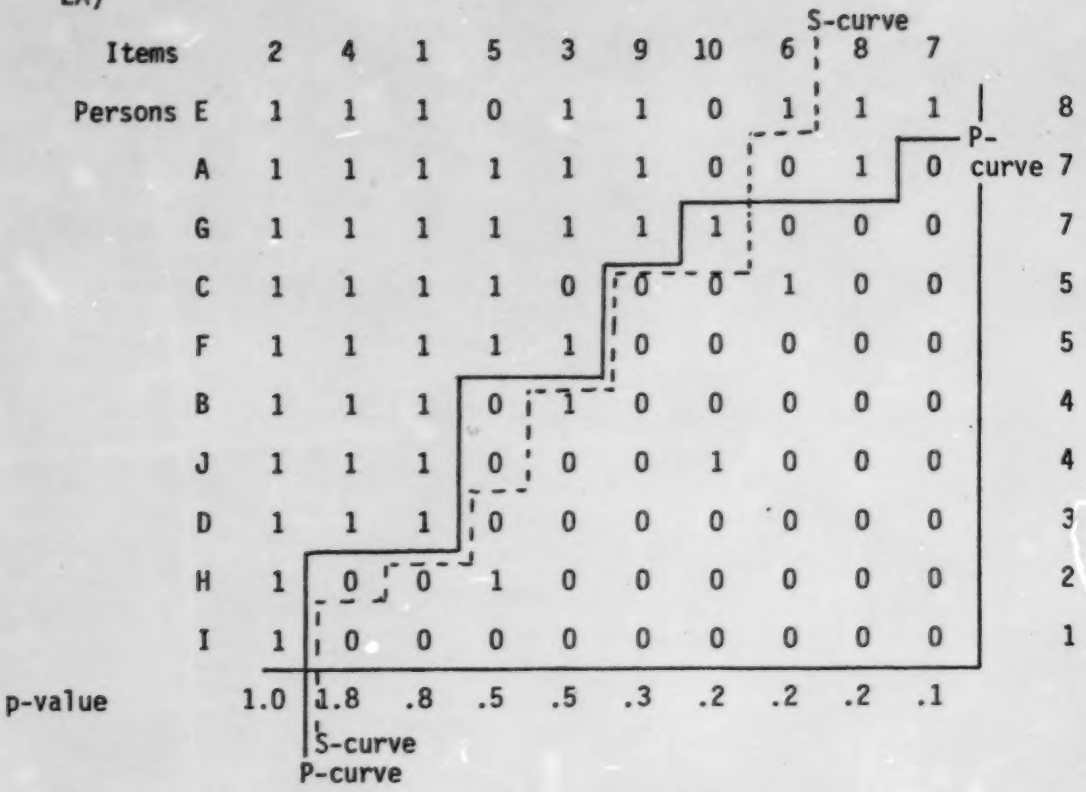
$\bar{p}$= .5500
s.d.= .3028

1B)

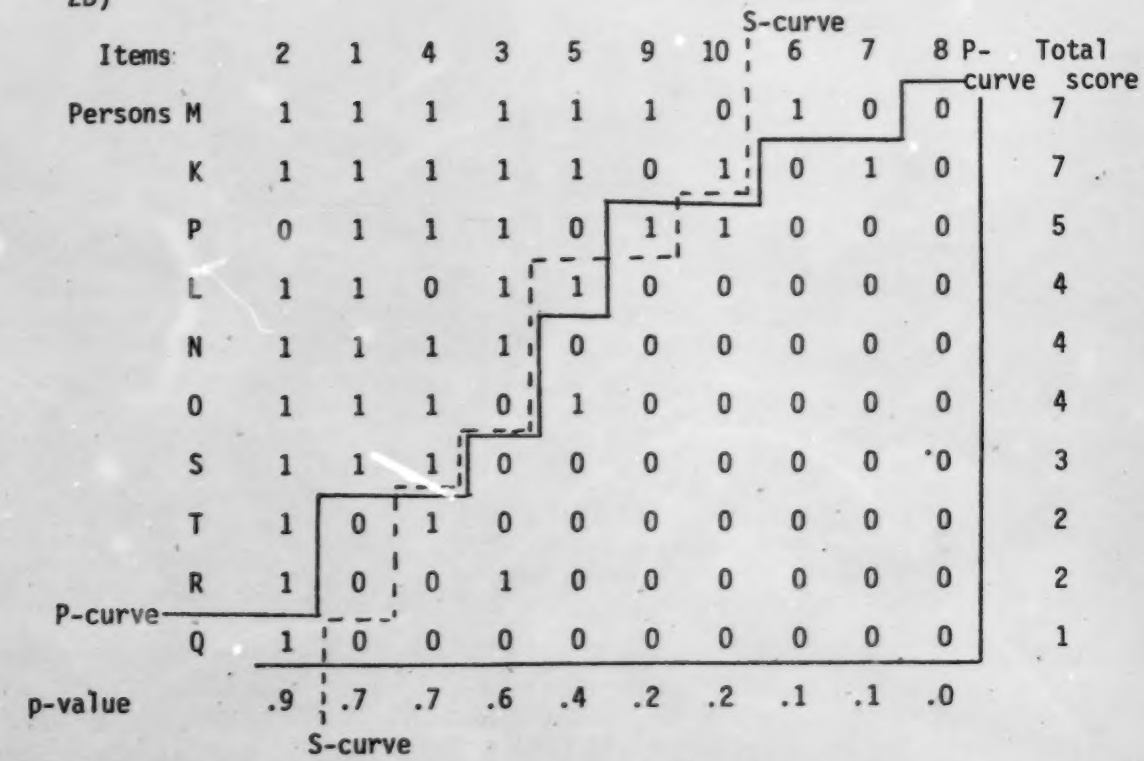
Items	1	2	3	4	5	6	7	8	9	10	Total score
Persons K	1	1	1	1	1	1	1	0	0	0	7
L	1	1	1	1	1	1	0	0	0	0	6
M	1	1	1	1	1	0	0	0	0	0	5
N	1	1	1	1	0	0	0	0	0	0	4
O	1	1	1	0	0	0	0	0	0	0	3
P	1	1	0	0	0	0	0	0	0	0	2
Q	1	0	0	0	0	0	0	0	0	0	1
R	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0
% correct	70	60	50	40	30	20	10	0	0	0	0

$\bar{p}$= .2800
s.d.= .2616

2A)


S-curve
Items 2 4 1 5 3 9 10 6 8 7
Persons E 1 1 1 0 1 1 0 1 1 1 8
A 1 1 1 1 1 1 0 0 1 0 7
G 1 1 1 1 1 1 1 0 0 0 7
C 1 1 1 1 0 0 0 1 0 0 5
F 1 1 1 1 1 0 0 0 0 0 5
B 1 1 1 0 1 0 0 0 0 0 4
J 1 1 1 0 0 0 1 0 0 0 4
D 1 1 1 0 0 0 0 0 0 0 3
H 1 0 0 1 0 0 0 0 0 0 2
I 1 0 0 0 0 0 0 0 0 0 1
P-curve
p-value 1.0 1.8 .8 .5 .5 .3 .2 .2 .2 .1
S-curve
P-curve

2B)


S-curve
Items 2 1 4 3 5 9 10 6 7 8 P- curve Total score
Persons M 1 1 1 1 1 1 0 1 0 0 7
K 1 1 1 1 1 0 1 0 1 0 7
P 0 1 1 1 0 1 1 0 0 0 5
L 1 1 0 1 1 0 0 0 0 0 4
N 1 1 1 1 0 0 0 0 0 0 4
O 1 1 1 0 1 0 0 0 0 0 4
S 1 1 1 0 0 0 0 0 0 0 3
T 1 0 1 0 0 0 0 0 0 0 2
R 1 0 0 1 0 0 0 0 0 0 2
P-curve
Q 1 0 0 0 0 0 0 0 0 0 1
p-value .9 .7 .7 .6 .4 .2 .2 .1 .1 .0
S-curve

3)

		Poorly-fitted item	Better-fitted item
Persons	U	0	1
	V	0	1
		---P-curve crosses here	
	W	1	1
	X	0	1
	Y	0	0 ---P-curve crosses here
	Z	0	0
	a	0	0
	b	1	1
	c	0	0

References

Anghoff, W. H. A technique for the investigation of cultural differences. Paper presented at the Annual Meeting of the American Psychological Association, Honolulu, 1972.

Berk, R. A. (Ed.) Handbook of methods for detecting test bias. Baltimore, Johns Hopkins University Press, in press.

Choppin, B. H. The correction for guessing on objective tests. Stockholm, International Association for the Evaluation of Educational Achievement, 1974.

Crowder, C. R. An investigation of item bias occurring at different ability levels for Anglo and Mexican-American students. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1979.

Comprehensive Test of Basic Skills (CTBS) Examiner's Manual and Espanol Examiner's Manual. Monterey, CTB/McGraw-Hill, 1974/1978.

Ebel, R. L. Constructing unbiased achievement tests. Paper presented at the National Institute of Education Conference on test bias, Baltimore, 1975.

Harnisch, D. L. & Linn, R. I. Analysis of item response patterns: consistency indices and their application to criterion-referenced tests. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Hayes, W. L. Statistics. New York, Holt, Reinhart and Winston, 1963.

Hudson, L. The relation of psychological test scores to academic bias. British Journal of Educational Psychology, 1963, 33, 120-131.

Hunter, J. E. A critical analysis of the use of item means and item test correlations to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education conference on test bias. Baltimore, 1975.

Jensen, A. R. Bias in mental testing. New York, Free Press, 1980.

Linn, R. L., Levine, M. V., Hastings, C. N. & Wardrop, J. L. Item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5, 159-173.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey, Lawrence Erlbaum, 1980.

Marascuilo, L. A. & Slaughter, R. E. Statistical procedures for analyzing item bias based on chi-square statistics. Journal of Educational Measurement, in press, 1981.

Merz, W. R. Methods of assessing bias and fairness in tests. ARC Technical Report #121-79, Sacramento, Applied Research Consultants, 1980. (ERIC Document Reproduction Service No. ED 198 145)

Petersen, N. S. Bias in the selection rule, bias in the test. In van der Kamp, L. J. T., Langerak, W. F. & de Gruiter, D. N. M. (Eds.). Psychometrics for educational debates. Chichester, G. B., John Wiley and Sons, 1980.

Powell, J. C. & Isbister, A. G. A comparison between right and wrong answers on a multiple choice test. Educational and Psychological Measurement, 1974, 34, 499-509.

Rudner, L. M., Geston, P. R., & Knight, D. L. Biased item detection techniques. Journal of Educational Statistics, 1980, 5, 213-233.

Sandoval, J. The WISC-R and internal evidence of test bias with minority groups. Journal of Consulting and Clinical Psychology, 1979, 47, 919-927.

Sato, T. The S-P chart and the caution index. NEC (Nippon Electric Company, Japan), Educational Informatics Bulletin, 1980.

Sato, T., Takeya, M., Kurata, M., Morimoto, Y., & Chimura, H. An instructional data analysis machine with a microprocessor --SPEEDY. NEC (Nippon Electric Company, Japan) Research and Development, 1981, No. 61, 55-63.

Subkoviak, M. J., Mack, J. S., & Ironson, G. H. Item bias detection procedures: empirical validation. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Tatsuoka, K. An approach to assessing the seriousness of error types. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Veale, J. R., & Foreman, D. I. Cultural variation in criterion referenced tests: A global item analysis. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.

Williams, R. L. Abuses and misuses in testing black children. Counseling Psychologist, 1971, 2, 62-77.

Bias in the Writing of Prose and Its Appraisal

David L. McArthur

Center for the Study of Evaluation

University of California Los Angeles

Supported by a grant from the National Institute of Education (NIE-G-80-0012). Appreciation is extended to Edys Quellmalz and Frank Capell for their roles in the planning of this study, and to Chi Ping Chou and Beverly Cabello for their roles in the analysis.

Abstract

Evidence from a variety of sources suggests that systematic differences can be found in the ratings given to student essays as a function not only of the student's skills but also of aspects of both the student's background and the background of the rater. Additionally, the nature of the prompt which provided the central theme of the essay might bias the outcome of the ratings of that essay. A study of ratings of fifth and sixth graders who wrote paragraph-long essays in response to two topics presented either in written or pictorial form is presented. Students were classified as Hispanic-surnamed or non-Hispanic-surnamed; two teachers, trained as raters using an objectively-based essay scoring scheme, represented an Hispanic cultural background and two a non-Hispanic background. Results from a blind rating of 100 complete essays show that several of the rating subscales were significantly influenced by an interaction between student ethnicity and rater ethnicity, and several subscales by rater ethnicity alone. Student ethnicity alone was not a significant main effect on any subscale. Prompt modality is significant for one subscale, and interacts with rater ethnicity on one other. The findings are interpreted as a direct indication of biased assessment.

The evaluation of writing of prose by schoolchildren poses special problems in relation to bias in educational appraisal. Many factors have long been known to have major influence on the prose writing performance of minority pupils. The literature on the issue of biases which occur in the judgment of students' written work is much smaller, and has proved much more contradictory. Are there specific aspects of non-native English writing style which undermine the usual procedures for judging writing performance? Do raters who match the cultural background of the writers whose work they judge arrive at different conclusions from raters who do not share the same background? In the present paper, the results of a research study involving both writers and readers from two different cultures are examined in an attempt to partition out the sources of systematic bias in the evaluation of writing.

Sources of bias: student variables

An overarching concern in the literature about bias in writing has been the isolating of sociocultural factors in students' backgrounds which contribute to differences in performance. A half-century ago, Caldwell and Mowry (1933) demonstrated that bilingual Hispanic children were at a disadvantage due to their use of language compared to their monolingual English-speaking counterparts when evaluated by the essays they wrote; on objective examinations the differences were not nearly as acute. Parallel findings emerge from the recent large-scale study by White and Thomas (1981), who combined files of data regarding entering students in the California State University and Colleges system to yield graphic comparisons of total scores for 5,246 whites, 585 blacks, 449 Mexican-Americans, and 617 Asian-

Americans on two English placement exams. The first was the CSUC's own English Placement Test; the second was the Test of Standard Written English from the College Entrance Examination Board. Although no statistical analyses were presented, profiles of the four distributions suggest that a dialect interference or second language interference hurt the overall performance of the three minority samples on both tests. Lay (1978) has shown that native-speaking Chinese students are at a disadvantage in writing English prose because of the wide differences in structure and phonology of English and Chinese. Rizzo and Villafane (1978) have shown that a similar explanation applies to native Spanish-speaking students.

Many investigators of language have shown that structural aspects of both oral and written language are significant in determining how children process the world around them. Moreover, many of the rules which govern functions of sending and receiving meaning using oral language are significantly different from those for written expression (Olson, 1977). For the non-native speaker of English the task of writing in English poses a particular problem because

> ...the surface structure of writing is an inadequate representation of both the sound structure of the target language and its meaning. Learning the underlying structure of the target language is as much of a bootstrap operation as the initial process of learning a mother tongue (Smith, 1975, p. 359).

One practical outcome of such a structural viewpoint is that students who fail to acquire skills in the underlying structure of English might do passably well with spoken English but probably will have great difficulty with writing. Another factor not to be dismissed lightly is the attitudinal

or psychological readiness of the student to orient positively to the task of acquiring skills in a new language (Cervantes, 1975; Lambert, Gardner, Barik, & Tunstall, 1963). Without the necessary motivation and appropriate learning context, students may be unable to let their knowledge of both the mother tongue and the new language interact to their advantage.

Sources of bias: evaluation variables

Beyond the issues of students' involvement in languages lies an important realm of educational and psychometric considerations having to do with the quantity and quality of appraisal. The nature of the task, how it is interpreted by both the student and the teacher, with what tools the students' writing is judged and by whom are all issues of import. In each of these lies the possibility of systematically different patterns of response for students from culturally or linguistically different groups. Each, then, may introduce its own bias into the evaluation of writing. The purpose of the writing task usually given to students in the classroom is to construct an essay following a particular prompt*. The teacher seeks a sufficient amount of this writing to rate the quality of the student's work. Exactly what elements are most important in that assessment of writing is often dependent upon the persons creating the scoring system. Freedman (1979) attempted to specify "definable parts" of student compositions which influenced teacher judgments. She concluded that content, organization,

*The prompt itself may contribute to systematic bias. Some students may not know what the prompt represents because they do not completely understand the vocabulary of the prompt in written form, or do not recognize the pictorial content (the palmtree vs. evergreen problem). Differences of an extreme nature are found in recognition of three dimensional objects in photographs or drawings between children of developed and underdeveloped countries. Subtler problems of prompt recognizability abound: one British picture recognition test for the primary grades depicts electrical items common in England but totally unknown in America.

and language mechanics were the most important factors, in that order. The effect of "weak" content was so powerful that it overshadowed teacher judgment in every other category. The interaction of content quality judgments with the quality of the writing prompt is one point where bias in assessment is possible.

The use of incompletely explicated scoring criteria introduces another potential for bias in writing studies. In Rhodes-Hoover and Politzer's (1974) study of teachers' attitudes toward Black rhetoric, teachers downgraded compositions in the category of "language mechanics" because students failed to use "superstandard" English. For example, if a student wrote, "I got there" as opposed to, "I reached my destination," the passage was considered too colloquial. Teachers not only gave their own interpretation of "usage" and "colloquial" but also imposed an undocumentable degree of severity in their judgment that may or may not have been intended by the scale.

In a study comparing the syntactical characteristics of Mexican and Anglo-American prose, Rodrigues (1978) asked educators whether they could detect "slight" or "noticeable" differences in the prose syntax of the two groups. At least 95% of these educators found some difference; 44% said they found "noticeable" differences. More Anglo-American educators found "noticeable" differences than did Mexican-American raters. Bikson (1977) conducted a study of differences in working lexicons of 72 lower grade and 72 upper grade White, Chicano and Black elementary school students. Results showed that ethnically diverse speakers made <u>different</u> kinds of lexical choices, particularly in the early grades. The differences between Anglo lexicon and either the Black or Chicano lexicon were greater than the differences between the two minority lexica. The study found

varying degrees of overlap between minority and Anglo word choice. The minority students used a wider range of vocabulary than the Anglo group, but this "broader" working vocabulary is not often valued by persons evaluating the speech of these students.

Differences in classification of lexical terms between different linguistic groups may have consequences for the selection of scoring criteria to evaluate the writing of these groups. If we take concept classification tasks to be analogous to organization tasks in the writing process, then the different strategies used to associate words may reflect different preferred methods of essay organization. If the scoring criteria implicitly prefer one type of content organization strategy, such preference could result in bias against those students who adopt alternative strategies. Two studies in particular seem to suggest that words are sorted by different ethnic groups into categories according to different classification strategies. Rissel (1978) studied the vocabulary-semantic relationship for monolingual English speakers, monolingual Spanish speakers, and Spanish/English bilinguals living in New York and Puerto Rico to determine the classification strategies of these groups. The study found that not only did the classification strategies vary by linguistic group but that there appeared to be a relationship between amount of language dominance and classification strategy. Spanish dominant bilinguals employed comparative criteria, whereas the more "balanced" bilinguals used comparative classification for Spanish words and inclusive classification for English. Stahl (1977) conducted a study comparing the "methods for arrangement" of content used by Israeli students of European or Arabic extraction. He found that those of European background tended to arrange the content in a hierarchical or inclusive manner, whereas those of Arabic background tended to use more associative or comparative techniques. An interesting aspect of his method was that

he gave higher points for hierarchical classification than for the use of comparative methods. In the assessment of writing this would appear to be deliberate introduction of biased criteria into the scoring process. Contrary results have been reported. In a study of syntactic patterns of lower and middle class Chicanos, Garcia (1975/76) concluded that the Chicanos used the same basic patterns found in American English, a conclusion also tendered by Rodrigues (1978). At the same time, however, Garcia cited research demonstrating differences in the morphological and phonological systems used by Chicanos and Anglos.

Recent informal evidence demonstrates the potency of systematic differences among raters of writing, Hartwell (1981) found that older, more experienced writers selected very different passages as exemplary of "professional writing" than did college freshmen. The differences appear to be consistent along a number of dimensions, including content, coherence, degree of complexity, and development. Differences in rating of a written essay may also be related to the rater's own level of cognitive complexity and integration (Sternglass, 1981). Rater background has been found to influence how scoring criteria are interpreted and applied. Follman and Anderson (1967) concluded that when raters shared similar backgrounds with regard to education and opinions about what constitutes good writing, they tended to agree on the ratings of essays more than raters who differed along these dimensions.

Whether writing is assessed through normative-holistic means or through differentiated judgments on dimensions of rhetorical quality, the scoring "instrument" will always be a human judge. Consequently, no question about

fairness, validity, or accuracy in writing assessment can be fully addressed without reference to possible errors in judgment. The intention of writing assessment is to generate information useful for diagnosis and/or remediation. When diagnostic utility is of interest several other issues are pertinent. Diagnosis implies performance profiles which in turn require a multidimensional view of the writing skill domain. Questions about skill profiles are connected intimately to rater behavior in assigning ratings. Scoring criteria are filtered through the expectancies of raters, and the halo effect inflates inter-subscale correlations (Jaeger and Freijo, 1975). The use of more and longer writing tasks only exacerbates this phenomenon.

Rating scales may interact. It is common for writing score profiles to include some attention to essay "mechanics;" variations along this dimension may influence ratings on other dimensions. Ratings assigned to a writing sample on such dimensions as "organization" or "use of supporting detail" may be assigned differentially depending on the quality of mechanics within the essay. For mechanically substandard work, this process might bring the assessment of other dimensions of writing quality into line with the rater's impression of mechanics while if level of mechanics is not so low as to call attention to itself, there may be minimal confounding. However, across a given set of papers the net effect would be correlated true and error components and concomitant inflation of inter-subscale correlations. In a multitrait-multimethod factor analytic formulation the expectation in general would be for negative correlations between mechanics "trait" factors and ratings "method" factors. Quellmalz and Capell (1979) used multitrait-multimethod confirmatory factor analyses to examine discriminant

validity of subscales generated by analytic scoring rubricsand the comparative information yield of alternative response modes for writing assessment (i.e., essay, paragraph and selected response). Their results indicated relatively high intercorrelations among subscale content factors, as well as a general tendency for the shorter assessment modes to generate less pure indicators of the subscale factors.

If non-native English speakers' English writing is easily distinguished from that of native speakers on the dimension of mechanics, and if such group differences contaminate other ratings assigned to non-native speakers, a straightforward form of bias may be present. Ratings on other dimensions will be systematically depressed, and the diagnostic utility of the writing appraisal undermined. The present study was conducted to evaluate such bias in the context of variations of ethnicity of both students and raters, and of prompts. Additionally the nature of the task presented to the students in order to get them to write an essay was varied systematically.

Method

Subjects

One hundred and thirty fifth and sixth graders from monolingual English classrooms in a Southern California school district of moderate size were involved in this study as a normal part of their classroom activities. These students were not members of bilingual programs although some were involved in remedial "pull-out" instruction. Of the 116 students who provided complete essays, half were Hispanic-surnamed. Raters were four teachers hired during school vacation, of whom two were Hispanic and two non-Hispanic. These raters were from different school districts and had no other contact of any kind with the students in this sample.

Instruments

The study used a standardized writing task with two topics, and a modified scoring rubric which has been shown to have acceptable validity and reliability (Quellmalz & Capell, 1979), explained shortly. The packet containing the essay writing task consisted of a face sheet for students name and date, followed by two prompts and two lined response pages, totalling five pieces of paper per handout. The prompts involved two topics, one a main street of a town and the other a robot. Order of presentation of the prompts, and whether the prompt was written or pictorial, was controlled for every participant. Written prompts involved five lines of typewritten text, while picture prompts involved a lead sentence and a full-page line drawing of the topic for children by a graduate student artist. In both situations, the text concluded with the request that the student write a paragraph about the topic presented. No other information was made available to the student.

The raters reviewed these essays using the Center for the Study of Evaluation's Factual Narrative scoring rubric, consisting of four primary subscales--General Impression, Focus and Organization, Support, and Grammar and Mechanics. Each of these was evaluated on a six-point scale, ranging from clear mastery of the assignment to clear failure. For each of the six values on each of the four scales, extensive guidelines for scoring were provided. General Impression ratings of the essay is formed by considering all aspects of the effectiveness of composition, including the remaining three rating criteria. The Focus and Organization subscale handles such issues as logical progression, transitions and topic development. The

Support subscale rates the use of specific supporting statements and details. The Grammar and Mechanics subscale is used to evaluate the essay's sentence construction, word usage, spelling and punctuation. As well as an overall rating from this last subscale, the extent of errors of each of the four areas of Mechanics noted above is rated separately. The instructions of the CSE scoring rubric are explicit that raters using factual scoring will likely find that some qualities of an essay cannot be considered separate from others, but it is also quite direct in indicating how any particular rating is to correspond to the annotation supplied in the guidelines,

Procedure

Each child received one essay packet, containing two essay prompts--one pictorial and the other written, and ruled pages for the child's essays. The package of essay prompts was administered in a single half-hour sitting by the children's classroom teachers, and essays were collected and sent directly for rating without further intervention in the classroom.

Each of the raters was given every essay packet in random order, but without the face sheet and thus without identification of the name or ethnic background of the student writers. Following five days of training and pilot testing on use of the CSE rating scales, the four raters completed scoring of the 116 essay packages which were complete and legible over a seven day period. The resulting 32 ratings for each essay (four raters x eight subscales) were then analyzed by a three factor analysis of variance (student ethnicity x rater ethnicity x prompt modality) with repeats on the second two factors (Winer, 1962) separately for each subscale. Also

collected from school district records were subtest totals on the Comprehensive Test of Basic Skills (CTBS), adminstered as part of the regular testing program by the s ool district, for all students involved in the study. These scores allowed the investigation of possible relationships between the measures of writing capability and four aspects of students' intellectual capacity--vocabulary, passage comprehension, language mechanics, and expression.

Results and Discussion

Only essays with complete ratings were considered in the analysis; complete data was available for the four primary subscales for 100 essays, and for the four detail subscales for 74 essays. Average rater agreement across all subscales was high for the two Hispanic raters (92.15%) and moderately good for the non-Hispanic raters (85.46%). When all four raters were compared, average agreement on the subscales was good (81.15%). These values were considered as acceptable evidence that the training of the essay raters had been satisfactory. To minimize potential confounding from differences between the two topics, all scores were then standardized within topic before further analysis.

On the General Impression subscale, the interaction between student ethnicity (Hispanic or non-Hispanic) and rater ethnicity (Hispanic or non-Hispanic) was significant ($F_{1,98}=6.51$, MSerror = 13.37, $p<.01$). While the non-Hispanic student essays received about the same General Impression scores from Hispanic raters as the Hispanic student essays, the non-Hispanic raters significantly favored the non-Hispanic student essays. No other

main effect or interaction was significant for this subscale. The interaction between student ethnicity and rater ethnicity was also found on the Support subscale ($F_{1,98}$=4.02, MSerror = 31.48, p<.05), and on the Mechanics subscale ($F_{1,98}$= 7.18, MSerror = 36.42, p<.01). On the Support subscale, the non-Hispanic student essays were again significantly favored by the non-Hispanic raters. However, on the Mechanics subscale, the non-Hispanic raters judged both student groups alike while the Hispanic raters gave the essays of the non-Hispanic students significantly lower scores.

For the Focus subscale, a main effect of rater ethnicity ($F_{1,98}$= 11.82, MSerror = 16,62, p<.001) and an interaction between rater ethnicity and prompt mode (picture prompt or written prompt) ($F_{1,98}$ = 6.41, MSerror = 19.01, p<.01) were found. In addition to the rater ethnicity by student ethnicity interactions, the Support subscale yielded only a main effect of prompt modality ($F_{1,98}$ = 10.43, MSerror = 68.17, p<.001), and the Mechanics subscale yielded only a main effect of rater ethnicity ($F_{1,98}$ = 13.45, MSerror = 36.42, p<.001). On the detail subscales of Mechanics, only one effect emerged as significant: rater ethnicity as a factor in Usage ratings ($F_{1,73}$ = 41.01, MSerror 47.01, p<.001). No other detail subscale showed any

---

Insert Table 1 about here

---

significant main effect or interaction. Table 1 summarizes the findings across the four primary and the usage detail subscales by main effect and interactions, and the results of post-hoc analyses.

When performance scores on the CTBS were compared, neither the Hispanic nor non-Hispanic students emerged as significantly more capable on any subscale than the others. The results of the correlational study between

student essay ratings and the four selected scale scores from the CTBS can be summarized rapidly. Not a single significant correlation appeared between any rating subscale and any CTBS scale for this sample. Thus there appears to be no intrinsically overlapping information between writing performance as judged on CSE's Factual Narrative rubric and a sample of academic performance as judged on a multiple-choice examination.

The most important finding, repeated across three of the subscales, is that the student ethnicity and rater ethnicity factors interact frequently and substantively in the appraisal of students' written essays. Additionally, rater ethnicity alone is also a significant factor in the ratings. These results point to three conclusions. First, the evaluation of prose writing seems to be systematically affected by factors which reflect different cultural backgrounds. It is important to note that this effect does not emerge when essays are grouped solely by student ethnicity; rather, the students of one or the other backgrounds were often judged differently by raters who share that background than by raters who do not. Second, these factors include (but are not limited to) a match or mismatch between raters' and writers' preferred language styles, and to some extent the nature of the stimulus used to initiate the writing sample. Note, however, that the three factor interaction between student ethnicity, rater ethnicity, and type of prompt was not observed for any of the subscales used. Third, the phenomenon of systematic matching or mismatching of preferences and styles occurs despite the fact that the evaluative scheme used is one with a high degree of objectivity, which would be expected to minimize such matching relative to more subjective rating scheme. The nature

of the judgment task is referenced point for point by the CSE scoring rubric and thus no scale-free or endpoint-only continuum judgments were involved. Additionally, because raters were blind not only to the names and ethnicities of the essay writers, but to the study's hypotheses and the proportional representation of ethnicities within the sample, whatever matching occured most likely stems from recognition of and preference for certain subtle aspects of writing styles.

Some limitations of the present study deserve attention. There are many possible secondary analyses of writing style, process and content which have not been pursued here. No information about essay complexity or other linguistic patterns is available from the present analysis. How creative, stereotyped, or bizarre the particular essay is goes unremarked in the CSE scoring system. The isolation of exact details within essay content or specific preferences of individual raters was not with the purview of this investigation. Moreover, there is a small possibility that systematic differences in handwriting mastery contributed to the recognizability of student ethnicity and thus to the ratings given, but this was not examined directly. None of these considerations is seen as critical to the interpretation of the results presented above, in particular because the expected outcome of the analyses of variance in such instance would necessarily be a main effect due to student ethnicity alone or a three-way interaction between student ethnicity, rater ethnicity, and prompt modality. None of these effects emerged in the present study, but rather a pattern of findings which strongly suggests that some complex form of bias is at work.

Bias in judgment is a phenomenon which obtains under a variety of circumstances, some of which are intrinsic in the testing and evaluation process. The present findings indicate that extrinsic factors must also be considered. In the case of judgment of essays, where essay content has virtually limitless possibilities and appraisal of necessity is at least partially subjective, the opportunity for unintentional bias seems more likely. For the teacher or essay test administrator seeking to limit bias to the absolute minimum, the mandate is: those who are to perform the rating of the essays must be matched for appropriate backgrounds of the students who write the essays to be judged.

Rissel, D. Implications of differences in the organizations of a lexical domain in Spanish and English bilinguals. Bilingual Review, 1978, 3, 29-34.

Rizzo, B., & Villafane, S. Spanish language influences on written English. Journal of Basic Writing, 1978, 1, 62-71.

Rodrigues, R. A statistical study of the English syntax of bilingual Mexican-American and monolingual Anglo-American students. Bilingual Review, 1978, 3, 205-211.

Smith, F. Spoken and written language. In Lenneberg, E.H. & Lenneberg, E., (Eds.) Foundations of language development, a multidisciplinary approach. New York, Academic Press, 1975.

Stahl, A. The structure of children's compositions: Developmental and ethnic differences. Research in the Teaching of English, 1977, 11, 156-163.

Sternglass, M.S. Assessing reading, writing, and reasoning. College English, 1981, 43, 269-275.

White, E.M. & Thomas, L.L. Racial minorities and writing skills assessment in the California State University and Colleges. College English, 1981, 43, 276-283.

Winer, B.J. Statistical principles in experimental design. New York, McGraw-Hill, 1962.

Table 1

Summary of statistically significant ($p<.05$) effects

Subscale:	General Impression	Focus and Organization	Support	Mechanics	Usage detail[1]
N=	100	100	100	100	74
Main Effects					
Student Ethnicity	--	--	--	--	--
Rater Ethnicity	--	*[2]	--	*[2]	*[2]
Prompt	--	--	*[3]	--	--
Interactions					
Student x Rater	*[4]	--	*[4]	*[5]	--
Student x Prompt	--	--	--	--	--
Rater x Prompt	--	*[6]	--	--	--
Student x Rater x Prompt	--	--	--	--	--

[1]Remaining detail subscales show no significant effects.

[2]Hispanic raters elevated relative to non-Hispanic raters.

[3]Picture prompt elevated relative to written prompt.

[4]Non-Hispanic raters + non-Hispanic student essays elevated relative to other combinations.

[5]Hispanic raters + non-Hispanic student essays depressed relative to other combinations.

[6]Non-Hispanic raters + Hispanic student essays elevated relative to other combinations.

Performance Patterns of Bilingual Children Tested in Both Languages

David L. McArthur

Center for the Study of Evaluation

University of California Los Angeles

Supported by a grant from the National Institute of Education

(NIE-G-80-0012)

Abstract

The testing of bilingual students poses particular problems for analyses of performance, item bias and test adequacy. When children are selected for their facility in two languages, and the same test is administered in both languages, a special arena is provided for the study of these problems. A widely-used test, the Comprehensive Test of Basic Skills, is available in both English and Spanish. The vocabulary subtest was administered to 1162 second-graders in bilingual education programs throughout the Southwest, as part of a larger study; 58 of those students received both versions of the test because they were deemed equally proficient in both languages. Results show that patterns of performance for these students differ markedly between the two versions, and suggest that the test differs in important dimensions even though the Spanish version is a rather faithful translation of the English original.

Severe problems confront the evaluation of bilingual program students from the standpoint of both individual performance measurement and the potential for bias in testing. Assessing the student in the majority language runs one set of risks; assessing in the native tongue runs another. The number of studies which have successfully assessed a single skill in two languages for the same individuals is exceedingly small (Duran, 1980). Resolution of these problems is not aided by the current controversy surrounding both the definition and measurement of bilingualism itself (De Avila, 1978). Moreover, thoroughly contradictory findings emerge from studies of the acquisition of French by native English-speaking children in Canada (Lambert & Tucker, 1972), of Swedish by native Finnish-speaking children in Scandinavia (Skutnabb-Kangas & Toukomaa, 1976), and of English by native Spanish-speaking children in the U.S. (Fischer & Cabello, 1978). The integration of such differences may rest in part on linguistic, developmental, and/or sociocultural interpretations (Troike, 1978); a practical level of shared bilingualism or dominance of one language over the other in the community may also play a strong role (Laosa, 1975). Finnish-speaking children from the populous southern districts find, and potentially model, both Finnish and Swedish in almost every shop window, while the politics of separatism are explicit in Quebec and de facto in many areas of the American Southwest, so children from these regions may encounter the second language with mixed emotions. Assessing even a relatively simple arena like vocabulary skills becomes multiply compounded when dealing with students who must cope with two languages.

Measuring the skills of bilingual program students necessarily also means assessing whether tests developed for the monolingual-English student are appropriate for making decisions about bilingual or limited-English proficient students characteristically found in such programs, and of minority groups who tend to be overrepresented there. Some educators believe that many tests are intrinsically unfair to minorities because the values they reflect are those of the majority only (Cervantes, 1975). Others, however, hold that tests of culturally defined content and vocabulary are not biased because achievement itself is language and culture specific (Ebel, 1975). But the impetus for testing continues:

> The problem how becomes not whether to test bilingual students, but rather how to do it in a manner that accurately assesses their specific abilities and in a manner that does not create a bias either against them or in favor of them (Cooper, 1978, p. 2, italics original).

We turn attention specifically to assessment in Spanish-English bilingual programs at the primary level, and encounter two factors which strongly mitigate against simple effective solutions to the problems noted above. The first is that exceedingly few instruments are available at present which are both culturally appropriate and technically sound for this purpose. "The problems are particularly acute with respect to English language measures, but are often equally pervasive in instruments that are simply translations from English language versions" (Burry, 1979, p.8). The second is that English-language instruction in reading, listening comprehension and vocabulary may be intrinsically more difficult for native Spanish-speaking children than for their native English speaking counterparts because of the increased rhythmic and phonological complexity of

English. Fundamental linguistic skills for understanding Spanish are frequently inadequate for comprehending English. Even a relatively simple phrase like "I c'n take it home fer ya" (/ə̆ ˌkn̥tḗykɪthṓwmfɚ́yɨ̆/ for the English listener) is likely to be heard by the native Spanish-speaking child as /'aintekrómfia+/, resulting in the obliteration of six out of seven words in the sentence (Matluck & Mace, 1972). The quantity of purely linguistic differences between Spanish and English suggest that the Spanish-speaking child is at no small disadvantage; especially in the primary grades, appropriate language skills testing must not ignore such difficulties.

The Comprehensive Test of Basic Skills/Spanish (1974/1978), is in large measure a direct translation of its English counterpart, which has been widely used as a primary skills evaluation tool. The CTBS/S has been presented as a major attempt to meet the needs of native Spanish-speaking children (Finch, 1979). With such a test, the teacher can select the language appropriate for a child with some assurance that the instrument is valid, reliable and unbiased (Hoepfner & Christen, 1979). Thus, the CTBS and CTBS/S should provide a good vehicle to examine individual performance patterns in either language for students in bilingual programs. However, recent evidence based on the performance of English- and Spanish-speaking pupils suggests that the tests contain multiple sources of bias (McArthur, 1981), so a particularly interesting situation for research obtains when both versions of the CTBS are administered to the same children. That is, if a group of children who possess similar levels of knowledge in both English and Spanish are tested on both instruments, will individual performances be the same across the two? Will the results of such dual-

language testing reflect patterns which can be interpreted as the direct result of item bias? Will direct translation hold up as a viable strategy for fair testing of primary pupils in Spanish as well as in English?

Methods

Subjects

As part of a larger study (CSE, 1979), almost 1200 children in bilingual education programs in 26 school districts spread over five southwestern states were administered a series of educational achievement tests by their teachers. Programs were designed to provide instruction in reading and mathematics at the upper primary level. Teacher reports from these programs indicate that the time spent using Spanish as the language of instruction was approximately equal to the time spent using English. Ninety-three percent of the program teachers had earned at least a BA or BS; 94% were full-time employees of the school district, and 88% had prior experience in bilingual education. Assignment of students to these special programs relied primarily on teacher evaluations and language dominence tests. Achievement tests were infrequently used to determine remediation placement, and intelligence test scores were generally excluded altogether from placement considerations. Thus the programs represented a major effort, competently staffed, to provide special attention in a bilingual setting to student educational needs. Most of the students were rated by their teachers as having some skills in both English and Spanish. Overall only one child in ten from these classes was considered monolingual Spanish while only one in nine was rated as monolingual English.

Instruments

While a large number of instruments were used in the investigation of programs, only the CTBS is of concern in the present study. It was selected because test content between the two language versions is virtually identical. The CTBS-Spanish was the first test by a major publisher to be subjected to a four-step editorial procedure designed to reduce bias; included were studies of content validity, application of editorial guidelines in item construction, reviews for bias, and separate ethnic group pilot studies. The developers of the Spanish-language version tried to keep the test content and measurement features intact, thus building a test which was similar in rationale, administration and interpretation to its parent version in English. What differences exist are the result primarily of problems of literal translation.

The children in the study were given a large number of standardized tests of achievement during the course of the regular school year by their teachers. With regard to the CTBS, the important instruction made to teachers was that they decide in advance on an individual basis whether each child would receive the English-language or Spanish-language version of the test. This decision was left totally to the discretion and best judgment of the classroom teachers. A total of 1162 completed test forms were returned, 814 in English and 348 in Spanish. Fifty-eight students in the sample were found to have been tested in both languages; that is, one student in every nineteen was given both forms of the test because the teachers felt unable to distinguish in advance which language these students should be tested in. No evidence is available to suggest that any selection

bias or other external circumstance might have contributed to obtaining this sample. Order of adminstration was apparently random. For purposes of this report, only the Vocabulary subscale of test level C, consisting of 33 items selected in response to the teacher's verbal directions, is considered.

Methods of analysis

Two techniques for analysis of response patterns were utilized in this study. The first relies on the work of Sato (1980) and colleagues in Japan; they have generated a systematic method of appraisal of test performance based on the S-P (Student-Problem) Chart, a matrix of right and wrong answers, coded 1 or 0, for each respondent for each item. The N x n matrix has the additional characteristics that students have been sorted by descending total score and items have been sorted by increasing difficulty. Thus the top row of the S-P Chart is a representation of the pattern of correct and incorrect responses to this sample of items by the most capable student in the group, the bottom row by the least capable. The left-hand column shows the pattern of responses to the easiest item in the set of items, and right-hand column shows the most difficult. From this matrix are generated two statistics, one related to the group pattern for the group as a whole, the other related to individual performance vis-a-vis both the group and the configuration of itmes, for each individual. The first is an "index of discrepancy," D*, which ranges from 0.00 for a matrix of perfect symmetry between student capabilities and item difficulties, to 1.00 for a matrix representing exclusively random responding.[1] The second is a "caution index," c_i, which ranges from 0.00 for an individual whose response pattern is perfeclty fitted to that reflected in the order of item

difficulties as determined by the group, to 1.00 for an individual whose pattern of responses is total antithetical to the order of item difficulties, and thus is quite unlike the representative average respondent in the group.[2]

The second analytic tool used in this study is a statistic from Goodman and Kruskal called lambda, which has been applied elsewhere to the detection of differences in response patterns in testing (Veale & Forman, 1976). Here the focus is on differences between groups in the attractiveness of incorrect responses within the multiple-choice format of one correct and three incorrect responses per item. Lambda is an index of the pattern of choice for the incorrect responses. If the value of lambda is 0.00, the two groups use about the same pattern of selection of the incorrect responses. As the value increases, one group is using a different strategy for selection of incorrect responses than the other. The computation of lambda is independent of the actual proportions within each group who select the correct response to the item. In this paper, values of lambda above .10 are considered noteworthy.[3]

Details of the computation and use of these approaches in the context of testing and item bias detection research have been set out elsewhere (McArthur, 1981). The usual test-retest and reliability statistics are not appropriate here, because of the attention to deciphering specific performance patterns rather than whole-group performance.

Hypotheses

Because of process of respondent selection, specific hypotheses about their performance on the English-language and Spanish-language versions of the Vocabulary subtest were, first, that the achieved scores between tests

would be perfectly correlated. Additionally, the S-P charts for the two versions would be similar, as shown by equal indices of discrepancy, D*. At the level of the individual respondent, it was hypothesized that the achieved total score in English would equal the achieved total in Spanish, and that the caution index generated for each individual in the English-lanugage S-P chart would be equal to the caution index obtained by the same individual from the Spanish-language S-P chart.

Results

Total scores on the English-language Vocabulary subtest averaged 75.34% correct with a range of 6 - 33. On the Spanish-language version, the average was 37.56% correct with a range of 4 - 25. The total scores are significantly ($p<.05$) correlated, $r = .48$. Median improvement from Spanish to English is 13 answers correct. Only three of the 58 participants did not show improvement in their total scores from Spanish to English.

Two of the 33 items yielded higher percentages of correct response in the Spanish-language version than in the English. For the remainder of the items, students were able to select the correct response less frequently in the Spanish-language version, often by substantial margins. The ratio of Spanish correct to English correct for each item is shown in the first column of Table 1. The consistency with which students picked the correct

---

Insert Table 1 about here

---

answer in both languages ranged form moderately high (65% of the respondents chose the correct answer to item 8 in both language to very low (only 7% chose the correct response to item 31 in both languages). The consistency of selection of incorrect responses was generally extremely low,

reaching 14% for items 24 and 31. The proportions of joint correct and joint incorrect proportions are shown in columns 2 and 3 of Table 1.

Those incorrect answers to items which garnered at least 10% more responses than the next most frequently chosen incorrect response were termed "popular distractors." Three popular distractor items were found in the English-language version, while twelve were found in the Spanish. The average percentage of respondents who chose the correct answer to an item in English but were swayed to choose the popular distractor (incorrect) response to that same item in Spanish was 35%. The reverse, choosing a popular distractor response in English although selecting the correct response to that same item in Spanish was 30%. Whether a specific item contained a popular distractor, and if so the percentage of respondents correct on the same item in the other language but who choose that popular distractor, is indicated in the next four columns of Table 1.

The data to this point quite clearly indicate that the Spanish-language version of the CTBS presented a far more difficult task for these respondents than did the English-language version. Only infrequently did any vocabulary item from one version have both an equal percentage of incorrect selections. Examination of the S-P charts is necessary to show whether the difference in performance patterns is systematic.

The Spanish-language version generated a D* of .53, a relatively high level of randomness of responses, while the English-language version yielded a D* of .24, reflecting a much more orderly fit of subject capabilities to item difficulties. No exact test of significance exists for the size of, or differences between, D* values, but in this instance they represent

configurations of the S-P charts which are distinctly different visually. The difference is supported by reference to the caution indices which for individual respondents to the English-language version averaged .17, but to the Spanish-language version .25. That is, on average the respondents were more consistent in selecting correct answers to easy items and incorrect answers to difficulty items in the English-language version. In fact, the number of respondents with caution indices of 0.00 is much higher in English. Of particular interest is that the correlation between the two indices computed across the 58 participants is nonsignificant. Changes in caution indices from one language version to the other are uncorrelated.

The computation of lambda, which details differences in selection patterns for wrong answers, showed that twelve out of 33 items had large discrepancies in the obtained configuration. That is, for a large number of items, the respondents shifted their choice from one incorrect answer to another across language versions, rather than picking the same incorrect responses on both occasions. The last column of Table 1 indicates those items with such shifts in incorrect answers.

Discussion

The findings of this study in general comport with earlier research on the CTBS in English and Spanish using independent groups of bilingual program respondents (McArthur, 1981). The distributions of total subscale scores, the higher D* indices for the Spanish-language version, the number of popular distractors and of lambda values exceeding .10 are all similar. That the two versions of the test do not produce equal outcomes even when the actual respondents are identical seems clear from the present data.

If there was to have been equivalence of total subscale scores, of group or individual patterns of correct scores, or of selection of wrong answers between the English- and Spanish-language versions, the number of discrepancies emerging from the statistical computations would have been far smaller. In its present configuration, these data suggest that children do not show the same performance patterns in response to the two versions of the test. Review of data contained in Table 1 suggests that many of the items may be suspected of somehow biasing the choice of correct response, and that such potentially biasing items are more prevalent in the Spanish-language version.

The relatively small number of individuals represented in this study makes these results necessarily tentative: they are presented neither as a representation of majority vs. minority responses to a specific test, nor as an indication in any way of a measure of true ability among bilingual program students. Rather, the unusual trial of a purportedly decent test in two languages, a purportedly equal-ability student sample, and a classroom experience for that sample equally divided into the use of English and Spanish, demands thoughtful attention to the appraisal of testing. In the present investigation, one weakness is the absence of an independent and unambiguous assessment of bilingual capability, and the ensuing reliance on the accuracy of teacher selection of students equally competent in tow languages. DeAvila and Duncan (1978) have pointed out numerous shortcomings in teacher ratings of language competence. However, for this study, students were not drawn for their equally high abilities or for the purposes of assembling a homogeneous sampel, but only for their

language abilities to be equally high or low in both languages. Nothing is known about the relative levels of exposure to English or Spanish outside the school, nor about the relative strengths and weaknesses of the texts in both languages used in the program. However, the teachers' close personal supervision of students and the even division between English and Spanish as the language of instruction in these programs suggest that the childrens' levels of readiness for vocabulary would be roughly similar. Another weakness is the relatively small number of items included in this investigation. However, the CTBS appears to represent the state of the art in English/Spanish testing of vocabulary skills at this level, and no other instrument is known to be a closer approximation to neutrality. The present results support the contention that the method of direct translation from English to Spanish for bilingual vocabulary testing may not be fully adequate for the needs of the bilingual program student.

References

Burry, J. Evaluation in bilingual education. Evaluation Comment, Center for the Study of Evaluation, Los Angeles, 1969, 6, 1-14.

Cervantes, R.A. Self-concept, locus of control, and achievement in Mexican-American pupils. Unpublished doctoral dissertation, Union Graduate School-West, San Francisco, 1975.

Cooper, E. Test selection in bilingual education evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Comprehensive Test of Basic Skills (CTBS) Examiner's Manual and Espanol Examiner's Manual. Monterey, CTB/McGraw-Hill, 1974/1978.

Center for the Study of Evaluation: Final Report: Basic Skills Learning Centers evaluation. Los Angeles, UCLA, 1979.

DeAvila, E. & Duncan, S.E. Definition and measurement, the east and west of bilingualism. Larkspur, California, DeAvila, Duncan & Associates, mimeo, 1978.

Duran, R.P. Bilinguals' skill in solving logical reasoning problems in two languages. Princeton: Educational Testing Service, 1980. (ERIC Document Reproduction Service No. ED 198 724).

Ebel, R.L. Constructing unbiased achievement tests. Paper presented at the National Institute of Education Conference on test bias, Baltimore, 1975.

Finch, F.L. At last: a Spanish version of CTBS. Paper presented to the California Association of Bilingual Educators, Fresno, 1979.

Fischer, K.B. & Cabello, B. Predicting student success following transition from bilingual programs. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, 1978.

Hoepfner, R. & Christen, F. Measures of academic growth. Santa Monica: System Development Corporation, mimeo, 1979.

Lambert, W.W. & Tucker, G.R. Bilingual education of children: the St. Lambert experience. Rowley, Massachusetts: Newbury House, 1972.

Laosa, L.M. Bilingualism in three United States Hispanic groups: contextual use of language by children and adults in their families. Journal of Educational Psychology, 1975, 67, 617-627.

Matluck, J.H. & Mace, B.J. Language characteristics of Mexican-American children: implications for assessment. Journal of School Psychology, 1973, 11, 365-386.

McArthur, D.L. Detection of item bias using analyses of response patterns. Los Angeles, UCLA, mimeo, 1981.

Sato, T. The S-P chart and the caution index. NEC (Nippon Electric Company, Japan), Educational Informatics Bulletin, 1980.

Skutnabb-Kangas, T. & Toukomaa, P. Teaching migrant children's mother tongues and learning the language of the host country in the context of the socio-cultural situation of the migrant family. Helsinki, Finnish National Commission for UNESCO, mimeo, 1976.

Troike, R.C. Research evidence for the effectiveness of bilingual education. Washington, D.C., National Clearinghouse for Bilingual Education, mimeo, 1978.

Veale, J.R. & Forman, D.I. Cultural variation in criterion referenced tests: a global item analyses. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.

Table 1

Summary of Findings for the CTBS and CTBS/S

item number	ratio of Spanish correct to English correct	% joint correct	% joint wrong	popular distractor		% who move from correct in one lang. to popular distractor in the other		lambda greater than .10
				English	Spanish	S to E	E to S	
1	.45	42	0	--	--	--	--	yes
2	.32	32	4	--	yes	--	32	--
3	.64	47	0	--	--	--	--	--
4	.63	60	2	--	--	--	--	--
5	.60	42	2	--	--	--	--	yes
6	.74	56	4	--	--	--	--	yes
7	.62	47	9	--	yes	--	17	--
8	.73	65	0	--	yes	--	12	--
9	.64	46	0	--	--	--	--	yes
10	.37	33	2	--	--	--	--	--
11	.29	25	0	--	--	--	--	--
12	.37	30	0	--	--	--	--	--
13	.25	19	4	--	--	--	--	--
14	.53	35	4	--	yes	--	31	yes
15	.46	30	9	--	--	--	--	--
16	.11	9	7	--	yes	--	41	yes
17	.13	9	4	--	yes	--	54	--
18	.67	54	4	--	--	--	--	--
19	.23	18	9	--	yes	--	40	--
20	.63	49	2	--	--	--	--	--
21	.22	16	5	--	--	--	--	--
22	1.06	37	9	--	--	--	--	yes
23	.77	44	5	--	--	--	--	--
24	.55	19	14	--	--	--	--	--
25	.53	23	5	--	--	--	--	--
26	.23	12	9	--	yes	--	43	yes
27	.51	33	4	--	--	--	--	yes
28	.37	12	4	--	--	--	--	--
29	.36	5	9	yes	yes	56	60	yes
30	.51	30	7	yes	yes	5	28	--
31	.40	7	14	yes	yes	30	48	yes
32	.70	42	9	--	yes	--	19	yes
33	1.41	12	11	--	--	--	--	--

Footnotes

1. $D^* = \dfrac{A\,(N,n,\bar{p})}{A_B\,(N,n,\bar{p})}$ where the numerator is a discrepancy between cumulative probability ogives obtained from the S-P chart, and the denominator is an analogous discrepancy as modeled by cumulative binomial distributions, both with the same number of cases, number of items, and average passing rate. (Sato, 1980).

2. $c_i = 1 - \dfrac{\mathrm{cov}(x_{ij},\, Y_j)}{\mathrm{cov}(u_{ij},\, Y_j)}$ where the numerator is the covariance over problems of the i-th student's score on the j-th problem with the number of students who correctly answer that j-th problem, and the denominator is the covariance over problems of the i-th hypothetical ideal student's score on the j-th problem with the number of students who correctly answer that j-th problem (Sato, 1980).

3. $\lambda = \dfrac{\Sigma\ \mathrm{max.f}_{jk} - \mathrm{max.f}_{.k}}{N - \mathrm{max.f}_{.k}}$ where $\mathrm{max.f}_{jk}$ is the larger frequency of the two groups for any single wrong choice, $\mathrm{max.f}_{.k}$ is the larger marginal frequency of the two groups acros- all wrong choices, and N is the total number of observations.